Biostatistics (2007), **8**, 2, *pp.* 285–296
doi:10.1093/biostatistics/kxl009
Advance Access publication on June 19, 2006

Bayesian detection of abnormal values in longitudinal biomarkers with an application to T/E ratio

PIERRE-EDOUARD SOTTAS*

Swiss Laboratory for Doping Analyses, Université de Lausanne, 1066 Lausanne, Switzerland
Chemin des Croisettes 22, 1066 Lausanne, Switzerland
pierre-edouard.sottas@hospvd.ch

NORBERT BAUME, CHRISTOPHE SAUDAN, CARINE SCHWEIZER

Swiss Laboratory for Doping Analyses, Université de Lausanne, 1066 Lausanne, Switzerland

MATTHIAS KAMBER

Swiss Federal Office for Sports BASPO, 2532 Macolin, Switzerland

MARTIAL SAUGY

Swiss Laboratory for Doping Analyses, Université de Lausanne, 1066 Lausanne, Switzerland

SUMMARY

We developed a test that compares sequential measurements of a biomarker against previous readings performed on the same individual. A probability mass function expresses prior information on interindividual variations of intraindividual parameters. Then, the model progressively integrates new readings to more accurately quantify the characteristics of the individual. This Bayesian framework generalizes the two main approaches currently used in forensic toxicology for the detection of abnormal values of a biomarker. The specificity is independent of the number n of previous test results, with a model that gradually evolves from population-derived limits when $n = 0$ to individual-based cutoff thresholds when n is large. We applied this model to detect abnormal values in an athlete's steroid profile characterized by the testosterone over epitestosterone (T/E) marker. A cross-validation procedure was used for the estimation of prior densities as well as model validation. The heightened sensitivity/specificity relation obtained on a large data set shows that longitudinal monitoring of an athlete's steroid profile may be used efficiently to detect the abuse of testosterone and its precursors in sports. Mild assumptions make the model interesting for other areas of forensic toxicology.

Keywords: Bayesian statistics; Biomarker; Doping; Longitudinal study; Steroids.

1. INTRODUCTION

As reflected in recent reviews (Timbrell, 1998; Watson and Mutti, 2004), research on biological markers is a fast-growing field for assessing evidence in biomedical toxicology. In forensic toxicology in particular,

*To whom correspondence should be addressed.

major goals are to develop and validate measurements of endogenous substances that may reveal the presence of toxic substances, drugs of abuse, and/or doping agents. The development and validation of biomarkers of response are either based on the statistical description of endogenous substances measured on a population or on a longitudinal evaluation of a series of repeated tests performed on the same individual. Longitudinal studies are particularly interesting in forensic toxicology when the biomarker has a significantly smaller intraindividual variability than interindividual variability. This is the case for several biomarkers currently used in antidoping investigations, such as indirect markers of blood doping (Malcovati *and others*, 2003; Sottas *and others*, 2006; Sharpe *and others*, 2006), and the testosterone over epitestosterone (T/E) ratio for the detection of the abuse of testosterone and its precursors (Donike *and others*, 1995). These biomarkers are all characterized by a small ratio of intra- to interindividual variation (Harris, 1974).

In contrast to the abundant number of statistical models that analyze serial biomarkers of disease (Slate and Turnbull, 2000), the development of reliable methods for the detection of abnormal variations of a longitudinal biomarker has remained astonishingly limited in forensic toxicology. To our knowledge, current methods employ either population-derived limits—to detect "absolute" abnormal values of the biomarker—or individual-based thresholds—to detect abnormal deviations relative to an individual baseline. There is no reason why a test cannot combine formally population-based information with individual-based data for better decision making. Failure to combine these two types of information may lead to a low sensitivity/specificity relation of the biomarker. For example, when a biological product such as the carbohydrate-deficient transferrine or a transminase (alanine transaminase (ALAT) and/or aspartate transaminase (ASAT)) is used as an indirect marker of chronic alcohol abuse (Musshoff and Dadrup, 1998), no effective method integrates previous readings for better decision making despite several measurements being available on the same individual. In the antidoping world, at least four readings are currently required by the World Anti Doping Agency (WADA) for the detection of abnormal variations of the T/E ratio, with no knowledge of the rate of false positives. Note that a precise valuation of the specificity is important in forensic toxicology, because a very low false-positive rate is demanded in order to prevent the accusation of an innocent individual. Finally, with biomarkers of blood doping, there is currently no procedure that has a specificity that does not depend on the number of test results. It is believed that "at least six samples for the baseline reading and possibly considerably more" are needed to derive cutoff thresholds that take into account the intraindividual variability of the biomarker (Sharpe *and others*, 2006).

Inspired by the recent success of Bayesian statistics to clinical trials (Berry, 2006), we propose a general approach for the detection of abnormal values of a biomarker. The idea is to use prior knowledge on interindividual variations of intraindividual parameters, and to progressively integrate previous readings to more accurately quantify the characteristics of the individual. Each new measurement of the biomarker is compared to a critical range. Before the first measurement on the individual, the critical range is derived from the population only. Then, this range progressively adapts itself as the number of readings performed on the same subject increases to finally characterize a particular individual only as the number of readings becomes very large. The rate of false positives does not vary with the number of previous test results.

We applied this algorithm for the detection of abnormal variations of the T/E urinary biomarker. A cross-validation procedure was used to validate the model developed using data obtained from two longitudinal studies conducted in our laboratory as well as data available in the literature. We also give two other examples in forensic toxicology to which the model can be applied.

2. MODEL

Let $[x_1, x_2, \ldots, x_n]$ represent n readings of the biomarker X measured on the same individual. A test result x is considered to be an outlier if it falls outside the $(1-\alpha)\%$ percentile range—the critical range—of the conditional probability distribution $p(x|x_1, x_2, \ldots, x_n)$, where α is the predefined proportion Type I errors, i.e. false positives. In practice, $[x_1, x_2, \ldots, x_n]$ often represents a temporal sequence of measurements of a biomarker and $x = x_{n+1}$, the next test result.

A review of several statistical models of intraindividual variation with different hypotheses can be found in the work by Harris (1976). Here, we require that

- the variable $X \approx N(\mu, \sigma)$ is normally distributed, with x_j independent of x_k for all $j, k = 1, \ldots, n+1$,
- the joint distribution $p(\mu, \sigma)$ representing the interindividual variability of the mean μ and standard deviation σ of X is known *a priori*, and
- analytical variability and/or measurement uncertainty do not change significantly in distribution between the evaluation of $p(\mu, \sigma)$ and the application of the model.

The joint distribution $p(\mu, \sigma)$ gives the probability of μ and σ prior to any measurement on an individual. This distribution is typically obtained in longitudinal studies involving a large number of control subjects.

The model does not require a precise knowledge of analytical variability and/or measurement uncertainty, but only that this variability does not change significantly between the preliminary construction of the prior distribution $p(\mu, \sigma)$ and the application of the model to new individuals. This assumption holds when test results are obtained following a standardized measurement procedure, which is often the case in forensic toxicology.

The usual solution is to compute a Z-score as a function of unbiased estimates of the mean $\hat{\mu}_n$ and variance $\hat{\sigma}_n$ of the sequence $[x_1, x_2, \ldots, x_n]$

$$Z_{n+1} = \frac{x_{n+1} - \hat{\mu}_n}{\hat{\sigma}_n} \tag{2.1}$$

and to compare Z_{n+1} with the Z-value corresponding to the desired rate of false positives α. This method does not use prior information on the parameters μ and σ, but estimates them directly from the measurements performed on the individual. As a consequence, the number of readings n should be large enough to obtain estimates that represent well the characteristics of the individual. In practice, the accuracy of Z_n should be close to the range defined by α. This makes this rule intractable in forensic toxicology, where α is small and the multiplication of the number of measurements unrealistic because of cost considerations.

A recent model introduced population-based information under the form of a variance σ^2_{uni} fixed for all subjects (Sharpe *and others*, 2006). Equation (2.1) becomes in that case

$$Z_{n+1} = \frac{x_{n+1} - \hat{\mu}_n}{\hat{\sigma}_{\text{uni}}\sqrt{1 + 1/n}}. \tag{2.2}$$

This second method uses prior knowledge on the variance σ^2 and, interestingly, can already be applied with $n = 1$. The dependence on n in the denominator makes the specificity of the model constant for any test result, under the assumption that all subjects have the same variance σ^2_{uni}. Unfortunately, this assumption is unrealistic for many biomarkers. As a result, the hypothesis of a null interindividual variance in the distribution $p(\sigma)$ induces a progressive loss of sensitivity with increasing n for subjects having a variance naturally smaller than σ^2_{uni}.

We propose an intermediate solution that formally takes into account previous readings to progressively learn the characteristics of the individual. The idea is to integrate the probability $N(\mu, \sigma)$ of a new outcome x_{n+1} over all possible values of the parameters μ and σ, given the previous readings $[x_1, x_2, \ldots, x_n]$. We have

$$p(x_{n+1}|x_1, x_2, \ldots, x_n) = \iint N(\mu, \sigma)|_{x_{n+1}} \cdot p(\mu, \sigma|x_1, x_2, \ldots, x_n)\mathrm{d}\mu\, \mathrm{d}\sigma, \tag{2.3}$$

where $p(\mu, \sigma | x_1, x_2, \ldots, x_n)$ can be interpreted as a probability mass function. Using Bayes' theorem and with the hypothesis that all readings are independent, we obtain

$$p(x_{n+1}|x_1, x_2, \ldots, x_n) = \iint \prod_{j=1}^{n+1} N(\mu, \sigma)|_{x_j} \cdot \frac{p(\mu, \sigma)}{p(x_1, x_2, \ldots, x_n)} \mathrm{d}\mu \, \mathrm{d}\sigma. \tag{2.4}$$

Interestingly, this formula uses the prior distribution $p(\mu, \sigma)$ of the mean and variance of the biomarker that characterizes the whole population. The process of passing from a prior distribution to a posterior distribution via the introduction of new data is typical in a Bayesian approach (Gelman *and others*, 2004). The posterior distribution is centered at a point that represents a compromise between the prior information and the data, and the compromise is controlled to a greater extent by the data as the number n of test results increases. When $n = 0$, the model detects abnormal "absolute" values of the biomarker, and gradually evolves to the detection of abnormal relative deviations from an individual's baseline as n increases.

In contrast to both models presented above, this model

- can be applied for any $n = 0, 1, \ldots, +\infty$,
- has a specificity that does not depend on the number n of test results, an important requirement in forensic toxicology, and
- makes the best—in a Bayesian sense—decision for each new test result x_{n+1} given the information available on the individual and the population.

Numerically, it is easier to calculate the joint distribution $p(x_1, x_2, \ldots, x_n, x)$ as a function of x

$$p(x_1, x_2, \ldots, x_n, x_{n+1}) = \iint \prod_{j=1}^{n+1} N(\mu, \sigma)|_{x_j} \cdot p(\mu, \sigma) \mathrm{d}\mu \, \mathrm{d}\sigma \tag{2.5}$$

and to compute the cumulative distribution function of the possible outcomes x, so as to find the limit values that correspond to the critical range defined by $(1-\alpha)\%$. The new test result x_{n+1} is then compared to these limit values.

As an example, we applied this method to gas chromatography/mass spectrometry (GC/MS) measurements of the T/E ratio. This marker is used to prove the administration of prohibited steroids in elite athletes, mainly the abuse of testosterone and its precursors. In that particular case, we are interested only in abnormal high values of the biomarker, so that the critical range is completely defined by a maximal bound for each new test result.

3. METHODS

All analyses have been carried out on Matlab version 6.1.0. with Statistics Toolbox version 3.0. A significance level of $p < 0.05$ was considered for all hypothesis tests. The numerical integration of (2.5) is performed via a simple quadrature rule. Since the integrands are represented by log-normal distributions (see below), we use a rectangle rule in which the intervals between interpolation points vary exponentially. With 100 interpolation points for each integral, the application of the model takes typically a few seconds for a sequence of length $n = 20$ on a standard personal computer.

T/E measurements have been obtained in two longitudinal studies. The first double-blind study involved 17 male amateur athletes followed for 1 month; two groups were orally administered either placebo (nine athletes, P group) or testosterone undecanoate (TU) (eight athletes, T group), 12 times during 1 month. Quantitative analysis of steroid glucuronide concentrations was then performed on urine by GC/MS. The study protocol as well as the analytical method is described in detail in the work of Baume *and others* (2006). A total of 175 readings were obtained from the nine control athletes, and 157 readings from the eight doped athletes. Of the 157 urines coming from doped athletes, 69 were collected more than 36 h after the last administration of TU. Sensitivity was estimated from 88 samples withdrawn within 36 h after TU administration.

The second study involved 11 male professional, top-level athletes. This study is a part of the project "Top Level Sport Without Doping" of the Swiss Federal Office for Sport with the aim of promoting doping-free sport and protecting the right of athletes to compete in a fair and ethical sport environment. The 11 athletes were extensively followed and tested over a period of 2 years. Each athlete was tested 17 times on average, returning only negative controls in both urine and blood matrices. About half of the samples were collected out-of-competition, while the others were collected prior to or during competition events. A statistical analysis of the hematologic parameters can be found in the work by Sottas *and others* (2006). In urine, 188 samples were collected and quantitative analysis of steroid concentrations carried out in our laboratory.

We also applied on these data the three-step procedure currently approved by WADA (2004). The first step is a screening procedure in which a population-derived threshold is fixed at four for the T/E ratio. Then, if the screening test returns a positive result, the urine sample is submitted to an isotope ratio mass spectrometry (IRMS) analysis. If the IRMS result is reported as inconclusive, further longitudinal study is performed. This longitudinal study is carried out by comparing the suspicious sample with a basal value calculated as the mean of at least three other readings. In this latter step, decision criteria are not well defined, leaving an open margin for interpretation. In particular, it is proposed that the coefficient of variation (CV) of individual T/E values should be smaller than 30% (a population-derived threshold), whereas some statistics (mean, standard deviation, and CV) calculated from the three basal values should be used for decision making (a decision rule based on individual values only). Here, we return a positive result if the CV calculated from all samples—the sample that define the baseline plus the suspicious sample—is larger than 30%.

4. Results

4.1 *Descriptive statistics*

We firstly verified that the period between two consecutive measurements of the T/E ratio was long enough so that the independence hypothesis required by the model holds. A correlation remains when the time between two consecutive measurements is too short. This can lead to an underestimation of the intraindividual variance σ^2 of the biomarker and, as a consequence, an underestimation of the rate of false positives α when applied to new subjects. In the first longitudinal study, each athlete was tested 19 times, on average, over 32 days. In particular, four spot urines were collected on day 24, one before treatment, and then after 4, 8, and 24 h. The aim was to see if a lag of several hours is long enough to assume independence of T/E measurements in the P group. The fact that the time between consecutive readings was not equal for all measurements makes this verification difficult. We nevertheless computed the sample autocorrelation function of the full series for the nine individuals who received placebo. All nine autocorrelation functions died out in the 95% confidence bounds defined by a moving average process of order 1 (MA(1)), except for one point for three subjects, a result which is not statistically significant. We also applied a nonparametric test of serial independence (Ghoudy *and others*, 2001), with no evidence found against serial independence on the nine full series. Finally, we verified if the variance calculated from the four values taken the same day is not significantly different from the variance calculated with the remaining values that have a longer time lag (more than 2 days in average). No evidence was found with a Kolmogorov–Smirnov (K-S) test on the two samples of size nine.

Similarly, no evidence against serial independence was found in the second study. This is not surprising since the average time between two consecutive measurements was much longer: one urine was collected every 43 days, on average, versus one urine every 40 h, on average, in the first study. We also found that the distribution of the variance in the second study was not significantly larger than the distribution of the variance in the first study (two-sample K-S test, $n_1 = 9$, $n_2 = 11$).

All these results suggest that the measurements are independent. However, to be on the safe side for future urines to test, we recommend that there must be at most one observation per day on an individual to let the autocorrelation function of the T/E marker fall close to zero. Composite normality of the intraindividual variation of the T/E ratio was verified on the data coming from the nine control athletes of the P group of the first study as well as from the 11 professional athletes participating in the second study. In the 20 results, we found one p-value less than 0.05 ($p = 0.03$) for one individual in the first study (Jarque–Bera test), indicating no evidence against normality given the total number of applications of the test.

4.2 *Estimation of prior probability distributions*

The application of the model depends on prior knowledge of the interindividual variability of the intraindividual mean μ and variance σ of the biomarker. The double integration of (2.5) can be computed with the 20 values of the mean μ and standard deviation σ obtained on the control subjects. Since a strong correlation was found between μ and σ ($R = 0.94$, $p < 1 \times 10^{-9}$, $n = 20$), we decided to model the joint distribution $p(\mu, \sigma)$ as follows:

$$p(\mu, \sigma) = \mu \cdot p(\mathrm{CV}) \cdot p(\mu), \tag{4.1}$$

where CV is given in percent. No correlation was found between μ and CV ($R < 1 \times 10^{-6}$, $p = 0.99$, $n = 20$), suggesting that the CV is indeed independent of the mean. To the 20 estimates of μ and CV in our possession, we added values obtained by other laboratories accredited by WADA. These values were found in the existing literature. Besides providing a larger number of data, this procedure allowed us to implicitly take into account a possible inter-laboratory bias in the prior distribution. Forty-four values were obtained from Donike *and others* (1994), five from Mareck-Engelke *and others* (1995), and seven from Ayotte *and others* (1996). A total of 76 estimates of the mean μ and CVs were therefore available from longitudinal studies.

It is possible to use directly these 76 sets of values to numerically compute the integration of (2.5). However, in order to facilitate the use of the method by other laboratories, we decided to represent the priors $p(\mathrm{CV})$ and $p(\mu)$ with parametric distributions.

For $p(\mathrm{CV})$, we found no evidence against log-normality on the 76 values (Jarque–Bera test of normality on the logarithm of the values). Such a distribution is plotted with a histogram of the 76 values in Figure 1. Maximum likelihood estimation of the geometric mean and geometric standard deviation gave values of 0.176 and 1.48, respectively. In the third step of the WADA procedure to prove the administration of a prohibited steroid, a limit at 30% has been suggested for the CV in male athletes. This limit gives a rate of 8.5% of false positives for this fitted distribution.

It is already known that the distribution of urinary T/E values is well approximated by a mixture of two log-normal distributions (Ayotte *and others*, 1996). A bimodal distribution was also found on 4885 values collected during routine controls from male athletes in our laboratory (see Figure 2). A fit with a mixture of two log-normal distributions on these 4885 values suggests that 13% of the values show a low basal urinary T/E ratio, whereas the remaining 87% show a larger value of T/E. This separation into two distinct peaks is also visible in the 76 mean values, since 13 values fall in the interval [0.08, 0.24] while the remaining 63 values are in [0.45, 6.3]. We therefore estimated the distribution $p(\mu)$ as a sum of two log-normal distributions on the mean of the 76 individual values of T/E. Maximum likelihood estimates

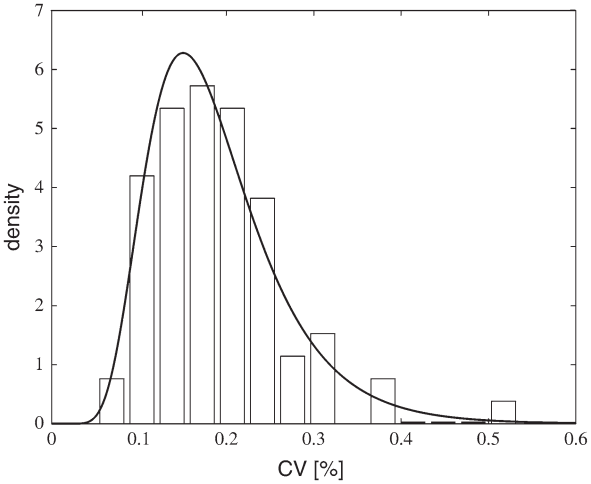

Fig. 1. Interindividual variability of CV of the T/E biomarker in male athletes. 30-bins histograms on 76 values, as well as the log-normal fit.

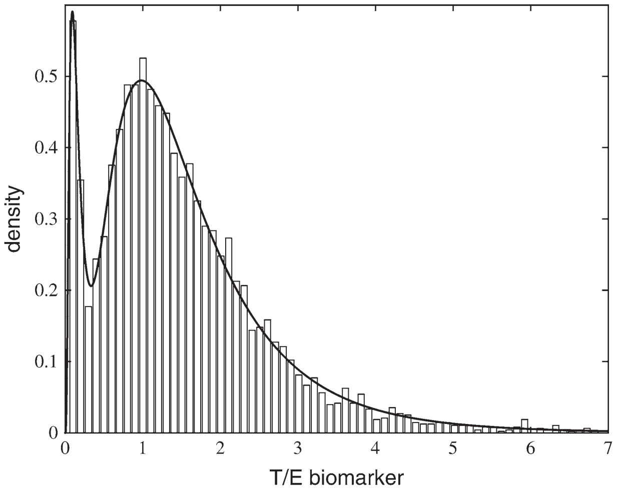

Fig. 2. T/E biomarker in male athletes: 200-bins histograms on 4885 values obtained from routine controls in our laboratory, as well as the fit found with a mixture of two log-normal functions on the mean of 76 series of T/E values.

of the geometric mean and geometric standard deviations are 0.141 and 1.43 for the left peak, 1.40 and 1.81 for the right peak, respectively.

4.3 *Comparison of methods*

Having models for the prior distributions $p(\mu)$ and $p(\mathrm{CV})$, we can now apply the test to any new value x_{n+1} given the series $[x_1, x_2, \ldots, x_n]$ of length $n = 0, 1, \ldots, +\infty$.

Table 1 presents the limits at which a given rate of false positives is found when $n = 0$. For example, the probability of finding a T/E ratio greater than 9.2 is equal to 1/1000. This value is a population-derived

Table 1. *Rate of false positives α of the T/E marker for different cutoff thresholds*

α	T/E limit
1/1000	9.2
1/100	5.6
1/10	2.9
1/26	4.0
1/130	6.0

threshold and can be compared with the cutoff thresholds currently adopted by WADA. The first step of the WADA strategy is a screening procedure in which a population-derived threshold has been fixed at 4, a limit that has been recently downgraded from a value previously fixed at 6. The corresponding rates of false positives are given in Table 1; in particular, 3.8% of false positives are found with a limit fixed at 4. For comparison purposes, 187 (3.8%) values of the 4885 samples of our internal database have a T/E greater than 4.

We applied the model to the 28 series collected during the two longitudinal studies. The 20 $(9 + 11)$ series from the control subjects are used to validate the specificity of the method, whereas the remaining eight series permit estimation of the sensitivity of the method in a 36-h time window. We used a "leave-one-athlete-out" cross-validation method to validate the model. In detail, for each application of the model on the data coming from one control subject, we reestimated the prior distributions $p(\mu)$ and $p(\mathrm{CV})$, excluding the data coming from this subject. This method mimics the application of the model as if all readings of the series were new values to test. We have recently shown the importance of validation techniques in forensic toxicology (Sottas *and others*, 2006). Cross-validation techniques are particularly important here to take into account sampling variation, since the cutoff limits are defined for a rate of false positives ($\alpha = 1/100$ and $\alpha = 1/1000$) that should not necessarily lead to false positives given the total number of control samples (363).

Figure 3 presents the T/E values measured on the first subject of the first study (blue curve), as well as the threshold limits found by our new model (red curve), a model that assumes a universal CV (yellow curve) and a traditional Z-score (green curve), with a rate of false positives $\alpha = 1/1000$. The universal CV was chosen so that $(1 - \alpha)$% of the subjects have a CV inferior to this value. This method has the advantage that the probability of finding a false positive is close to α when $n = 1$. The choice of a universal variance—instead of a universal CV—would have given even worse results, since the distribution of the mean is characterized by a large geometric standard deviation. Our model can be applied for any $n = 0, 1, \ldots, +\infty$, the second model for any $n = 1, \ldots, +\infty$, and the Z-score for any $n = 2, \ldots, +\infty$. For this subject ($\mu = 0.9$, CV $= 0.10$, $n = 20$), none of the three models produce any false positives; no values exceed the cutoff limits represented by the curves returned by the models. With $n = 1$, the model with a universal CV and our model have similar values, whereas the Z-score does not take into account the possibility that the initial values may be obtained from a subject who has a large variance (this case is depicted in Figure 4 where the Z-score returned two false positives when $n = 3$ and $n = 4$). With n large, our model is very close to the Z-score, whereas the model with a universal CV has higher cutoff limits. The latter model does not take into account the fact that this subject has a naturally low CV.

On all control series, with $\alpha = 1/1000$, the Z-score returned five false positives (1.6% of 323 tests), with four subjects having at least one value out of range (20% of 20 control subjects). The model with universal CV returned zero false positives on 343 tests. Our model returned zero false positives on 363 tests. The two-step procedure of the WADA (T/E > 4 and CV < 30%, IRMS excluded) returned zero false positives on the 20 series (303 values). With $\alpha = 1/100$, the Z-score returned 12 false positives

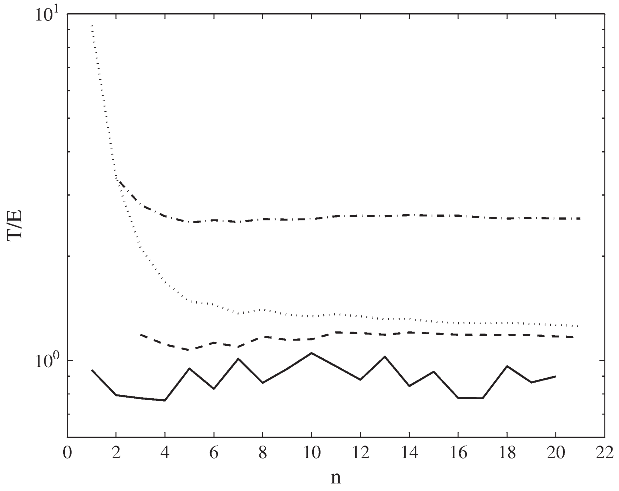

Fig. 3. A series of 20 longitudinal T/E values (solid line), with maximal cutoff limits found with our model (dotted line), a model that assumes an universal CV (dashed–dotted line) and a usual Z-score (dashed line), for a specificity of 0.999. All models produced no false positive on this series.

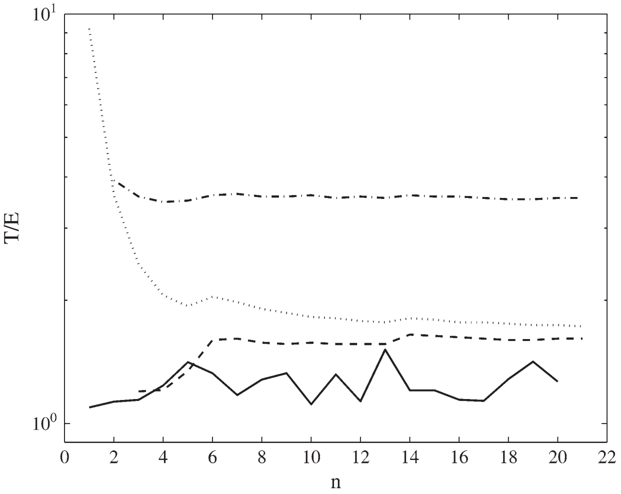

Fig. 4. A series of 20 longitudinal T/E values (solid line), with maximal cutoff thresholds found with our model (dotted line), a model that assumes an universal CV (dashed–dotted line) and a usual Z-score (dashed line), for a specificity of 0.999. The Z-score returned two false positives for $n = 3$ and $n = 4$.

(4% of 323), the model with universal CV no false positives (0% of 20 series, 343 values), our model two false positives (0.55% of 363 values).

If applied on the eight rough series of the eight subjects treated with TU, all three models returned at least one positive result with $\alpha = 1/1000$, except for one subject. This subject is a very fast metabolizer with a very low basal level ($\mu = 0.07$, CV $= 0.15$, $n = 7$). His T/E was not responsive to TU administration since none of his 11 T/E values exceeded 0.1. For example, a spot urine collected 4 h after the intake of TU returned a T/E of 0.087.

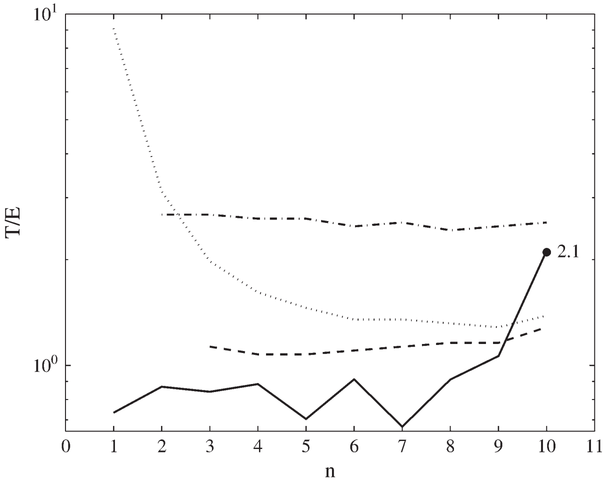

Fig. 5. A series of 10 longitudinal T/E values (solid line), with maximal cutoff thresholds found with our model (dotted line), a model that assumes an universal CV (dashed–dotted line) and a usual Z-score (dashed line), for a specificity of 0.999. The last value at 2.1 has been measured on a urine collected on a subject after oral administration of TU. A positive result is returned by our model and the Z-score, whereas a false negative is obtained with the model that assumes a universal CV.

Since the detection of oral administration of TU is highly dependent on the delay between TU intake and urine collection (Baume *and others*, 2006), we estimated the sensitivity of the models in a 36-h time window. Further, we assumed that the T/E ratio had returned to its basal level if the urine was collected more than 36 h after the last oral administration of TU. Even though this hypothesis may not hold for very slow metabolizers, this assumption may lead to an underestimation—and no overestimation—of the sensitivity in this time window. Thus, the 79 values from the urines collected at least 36 h after the last administration of TU were used to define eight individual basal series, and the sensitivity of the model was then calculated on the remaining 88 values on an individual basis. With $\alpha = 1/1000$, the Z-score returned 49 true positives (56%), the model with universal CV 33 true positives (37%), our model 43 true positives (49%), and the two-step procedure adopted by the WADA ($\mathrm{T/E} < 4$ and $\mathrm{CV} < 30\%$, IRMS excluded) 32 true positives (36%). Figure 5 shows a false negative obtained by the model with universal CV and the two-step WADA procedure. With $\alpha = 1/100$, the Z-score returned 51 true positives (58%), the model with universal CV 40 true positives (45%), and our model 50 true positives (57%).

5. DISCUSSION

We proposed a Bayesian approach to detect abnormal values of a biomarker. The two common methods currently used in forensic toxicology are retrieved as special cases when the number n of readings performed on the same subject is either equal to zero or very large. When $n = 0$, the cutoff limits are derived from the population, whereas the model detects abnormal deviations from an individual baseline when n is large. In contrast to current methods, the rate of false positives does not depend on the number of test results, an important requirement in forensic toxicology.

We applied this model to the detection of the abuse of testosterone with the T/E biomarker, and obtained a heightened sensitivity at a very low, fixed specificity. Our model may represent a useful discriminating tool since it generalizes the two screening procedures currently used to prove the administration of prohibited steroids.

The model is based on several hypotheses weak enough to make it functional in other areas of forensic toxicology. For example, it can be applied to the detection of abnormal deviations of biomarkers of alcohol abuse (Musshoff and Dadrup, 1998). Another biomarker of interest is an indirect marker based on the statistical classification of several hematologic parameters that we have recently proposed to deter blood doping in sports (Sottas *and others*, 2006). Currently based on population-derived limits only, this marker is fully compatible with the model presented here.

Indirect biomarkers become increasingly popular among forensic scientists for assessing evidence of drug abuse. The final choice of the rate of false positives, α, is highly dependent on the proportion of positives in the population—the prevalence—prior to the application of the test. An effect has been observed (e.g. a measurement of T/E), and we want to know if this is the result of a specified cause (e.g. the abuse of a prohibited steroid). From a Bayesian point of view, the probability of an effect given each cause is the likelihood ratio, and it is often believed that this ratio appropriately describes the value of the evidence in forensic sciences. To base a decision on the likelihood ratio alone means that one implicitly assumes a prevalence in the 50% range. This hidden assumption may be problematic with current multiplication of the number of indirect tests, since we cannot exclude that 100% of the positive results of the test are false positives. In our opinion, knowledge of the prevalence becomes imperative with the growing success of indirect markers in forensic toxicology. The estimation of the prevalence of doped athletes in a population can be performed after having tested a large number of athletes before an international competition for example. It is then possible to determine a cutoff limit, not as a function of the desired rate of false positives, but rather as a function of the probability that an athlete has doped when the test is positive, i.e. the positive predictive value (PPV). A PPV of 99% means that 99 doped athletes will be banned for one athlete wrongly accused on average. Knowledge of the prevalence makes possible to adjust the cutoff limits to a desired PPV. We believe that a strategy based on thresholds that adapt themselves to the proportion of doped athletes in the population is a good strategy, because the strategy is specifically designed to decrease this proportion. Knowledge of the prevalence, and subsequently the PPV, may allow antidoping authorities to develop more efficient strategies to deter doping in elite sports.

ACKNOWLEDGMENTS

This work was supported by the grant #320000-111771 from the Swiss National Science Foundation. We thank Dr Simone Bader and Dr Neil Robinson for their technical and scientific assistance. Both longitudinal studies were supported by grants from the Swiss Federal Office for Sports. We also thank the two anonymous reviewers for their constructive comments. *Conflict of Interest:* None declared.

REFERENCES

AYOTTE, C., GOUDREAULT, D. AND CHARLEBOIS, A. (1996). Testing for natural and synthetic anabolic agents in human urine. *Journal of Chromatography B* **687**, 3–25.

BAUME, N., SAUDAN, C., DESMARCHELIER, A., STRAHM, E., SOTTAS, P.-E., BAGUTTI, C., CAUDERAY, M., SCHUMACHER, Y. O., MANGIN, P. AND SAUGY, M. (2006). Use of isotope ratio mass spectrometry to detect doping with oral testosterone undecanoate: inter-individual variability of (13)C/(12)C ratio. *Steroids* **71**, 364–70.

BERRY, D. A. (2006). A guide to drug discovery: Bayesian clinical trials. *Nature Reviews Drug Discovery* **15**, 27–36.

DONIKE, M., MARECK-ENGELKE, U. AND RAUTH, S. (1995). Statistical evaluation of longitudinal studies, part 2: the usefulness of subject based reference ranges. *Proceedings of the 12th Cologne Workshop on Dope Analysis*. Koln: Sport und Buch Strausse Edition Sport. pp 157–65.

DONIKE, M., RAUTH, S. AND WOLANSKY, A. (1994). Evaluation of longitudinal studies, the determination of subject based reference ranges of the T/E ratio. *Proceedings of the 11th Cologne Workshop on Dope Analysis*. Koln: Sport und Buch Strausse Edition Sport. p 33.

GELMAN, A., CARLIN, J. B., STERN, H. S. AND RUBIN, D. B. (2004). Bayesian Data Analysis, 2nd edition. Boca Raton: Chapman & Hall.

GHOUDY, K., KULPERGER, R. J. AND RÉMILLARD, B. (2001). A non-parametric test of serial independence for time series and residuals. *Journal of Multivariate Analysis* **79**, 191–218.

HARRIS, E. K. (1974). Effects of intra- and interindividual variation on the appropriate use of normal ranges. *Clinical Chemistry* **20**, 1535–42.

HARRIS, E. K. (1976). Some theory of reference values. II. Comparison of some statistical models of intraindividual variation in blood constituents. *Clinical Chemistry* **22**, 1343–50.

MALCOVATI, L., PASCUTTO, C. AND CAZZOLA, M. (2003). Hematologic passport for athletes competing in endurance sports: a feasibility study. *Haematologica* **88**, 570–81.

MARECK-ENGELKE, U., GEYER, H. AND DONIKE, M. (1995). Stability of steroid profiles: the circadian rhythm of urinary ratios and excretion rates of endogenous steroids in male. *Proceedings of the 12th Cologne Workshop on Dope Analysis*. Koln: Sport und Buch Strausse Edition Sport. pp 121–33.

MUSSHOFF, F. AND DADRUP, T. (1998). Determination of biological markers for alcohol abuse. *Journal of Chromatography B Biomedical Science Applications* **21**, 245–64.

SHARPE, K., ASHENDEN, M. J. AND SCHUMACHER, Y. O. (2006). A third generation approach to detect erythopoetin abuse in athletes. *Haematologica* **91**, 356–63.

SLATE, E. H. AND TURNBULL, B. W. (2000). Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medecine* **19**, 617–37.

SOTTAS, P.-E., ROBINSON, N., GIRAUD, S., TARONI, F., KAMBER, M., MANGIN, P. AND SAUGY, M. (2006). Statistical classification of abnormal blood profiles in athletes. *The International Journal of Biostatistics* **2**, 3.

TIMBRELL, A. (1998). Biomarkers in toxicology. *Toxicology* **129**, 1–12.

WADA (2004). Reporting and evaluation guidance for testosterone, epitestosterone, T/E ratio and other endogenous steroids. *Technical Document*. http://www.wada-ama.org/rtecontent/document/end_steroids_aug_04.pdf.

WATSON, W. P. AND MUTTI, A. (2004). Role of biomarkers in monitoring exposures to chemicals: present position, future prospects. *Biomarkers* **9**, 211–42.

[*Received March 24, 2006; revised May 24, 2006; accepted for publication June 8, 2006*]